Smith RL.

Reinventing OCR for early printed books.

***CILIP Update* 2015, (February 2015), 40 - 41.**

Feature

Reinventing OCR for early printed books

Rachel Smith (r.l.smith@durham.ac.uk) **is Communications and Marketing Officer, Durham University Library and Heritage Collections.**

Digitisation methods for early printed Latin books have produced poor results, meaning that these texts are being left behind in the digital revolution. But free, open source Optical Character Recognition (OCR) software tailored to early Latin printed books being developed by Durham University could change this, reports **Rachel Smith.**

EARLY printed books in Latin have been carefully preserved over many centuries. From the Renaissance until well into the 19th century, Latin was the European language of every intellectual discourse – the natural sciences, mathematics, philosophy, theology, law, literary criticism, geography, archaeology and music.

Difficulties with digitising Latin

Most European historic cities, museums and universities hold early Latin books, as do many private collectors and learned societies. However, existing OCR packages are unable to digitise these texts effectively, and thus create an accurate searchable document.

Early Latin texts use non-standard typefaces, abbreviations and page layouts. Built to handle standard print in modern languages, current OCR software is unable to recognise historic characters, or indeed Latin morphology, syntax and vocabulary. Digitisation methods for early printed Latin books therefore produce very poor results.

Why is Latin being left behind?

Trials carried out using standard OCR packages have resulted in an accuracy of no more than 15 per cent when digitising early printed texts. With the software correctly recognising only one or two words out of ten, some have concluded that OCR is unsuitable for early books: 'it does not provide reliable results when applied to early modern publications'.[1]

The Text Creation Partnership (TCP), which creates electronic versions of early printed books in English for ProQuest's Early English Books Online (Eebo) and Gale Cengage's Eighteenth Century Collections Online (Ecco), share these concerns, and have resorted to hand-keying texts – an approach that will never work for early books in Latin, because there are many more of them, served by a smaller community of experts.

In a section of their website entitled 'Why OCR Won't Work',[2] an example shows a low-resolution image of a 16th century page, along with the results of running it through an unnamed piece of OCR software. The output is a random jumble of characters with no recognisable English word – hence the blunt and quite erroneous conclusion that 'OCR Won't Work'. In reality, what is demonstrated is that off-the-shelf commercial OCR packages don't support historic typefaces.

Commercial OCR companies have little interest and no expertise in tackling the problems posed by early printed Latin. The market is smaller than that for major modern languages, and the expertise required to produce this kind of software is greater: knowledge of Latin, early printing, machine-language training and visual machine training.

Current alternatives to OCR

At the moment, libraries with early modern Latin books in their holdings face four options:

- **Type texts by hand:** this produces high quality results, but manually keying in each word is slow and staff-intensive, and therefore very expensive.
- **Use current OCR packages:** hand-typing texts is not sustainable for major digitisation projects. Google Books and the Gallica project at the Bibliothèque Nationale de France have tried using existing OCR products, but the result is so low in accuracy as to be useless.
- **Don't provide a searchable document with metadata:** by providing digitised images without a transcript, libraries can avoid the costs of hand-typing and low quality OCR results. However, this means that full-text searching is not possible, resulting in reduced discoverability and

> Commercial OCR companies have little interest and no expertise in tackling the problems posed by early printed Latin.

accessibility – and impairing research.

● **Don't digitise:** unsatisfied with other available options, most collections of early Latin books do nothing at all, which is why these texts are being left behind in the digital revolution.

The background: adapting OCR for ancient Greek texts

Researchers working on the project *Living Poets: A New Approach to Ancient Poetry*, directed by Professor Barbara Graziosi and funded by the European Research Council, encountered a problem when studying ancient Greek works. They needed a way of transforming printed editions of ancient Greek texts into fully searchable digital documents, and had no ready way of doing so.

The team decided to investigate the possibilities of OCR using the open source program Tesseract, training it to recognise different ancient Greek character shapes, word lists and basic ancient Greek morphology and syntax.

The end result? An effective OCR system for ancient Greek, with an accuracy of 90 per cent, and 96 per cent for average quality page scans of printed volumes. This has been made available on a free, open source basis, as the *Living Poets* team aim to ensure that the results are 'freely available for study and improvement… maximis[ing] the utility and good the project does'.[3] Indeed, the project was so successful that one of the researchers was invited to work, for six months, for the Perseus Digital Library in Boston – the most important open-access library for ancient texts globally.

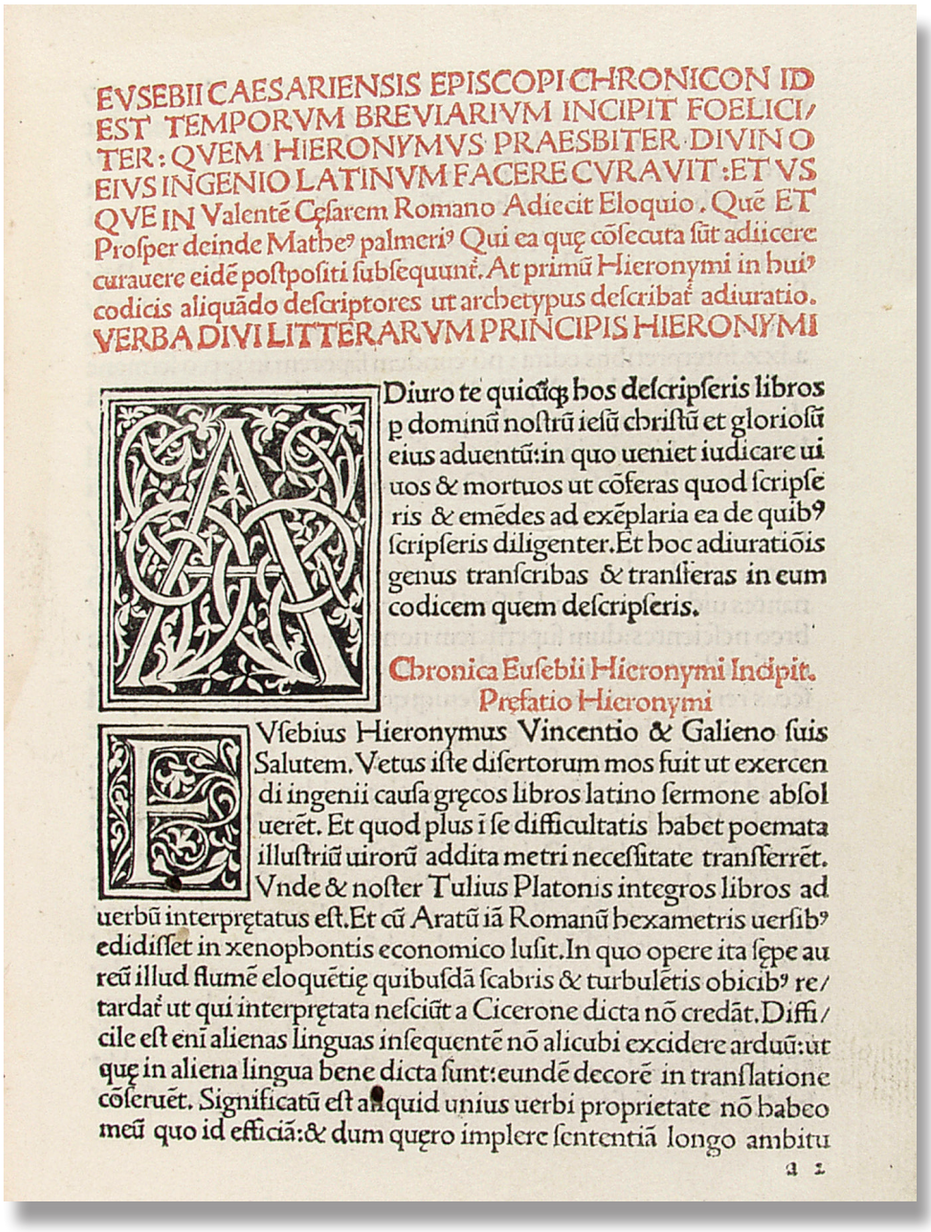
A viable solution for digitising early Latin

The *Living Poets* team plans to continue their work on developing OCR solutions for early texts and create free, open source software for digitising early printed texts in Latin, using their experience with ancient Greek. Given the volume and breadth of early Latin collections, this work will have the potential for much wider impact.

Some problems with tailoring OCR for early Latin are the same as for ancient Greek – and the team already know how to solve them. Tesseract can be programmed to recognise non-standard characters, and account for ancient morphology and syntax – for example by excluding impossible combinations of letters. Other challenges are new: an effective OCR system for Latin will need to handle a much wider range of typefaces, among other project-specific problems.

The modified system will be tested with libraries and digital publishers to ensure that it is user-friendly and meets the needs of those producing digital content.

When completed, it is predicted that the Latin OCR program will produce results with an accuracy of at least 80 per cent, with eight words out of 10 being correctly identified. This will reach about 95 per cent through programme customisation, tailored to specific collections, leaving only minimal manual checking for perfect texts.



EVSEBII CAESARIENSIS EPISCOPI CHRONICON ID EST TEMPORVM BREVIARIVM INCIPIT FOELICI/ TER: QVEM HIERONYMVS PRAESBITER DIVINO EIVS INGENIO LATINVM FACERE CVRAVIT: ET VS QVE IN Valentē Cęsarem Romano Adiecit Eloquio. Quē ET Prosper deinde Mathe⁹ palmeri⁹ Qui ea quę cōsecuta sūt adiicere curauere eidē postpositi subsequunt. At primū Hieronymi in hui⁹ codicis aliquādo descriptores ut archetypus describat adiuratio. VERBA DIVI LITTERARVM PRINCIPIS HIERONYMI

ADiuro te quicūq; hos descripseris libros p dominū nostrū iesū christū et glorioſū eius aduentū: in quo ueniet iudicare uiuos & mortuos ut cōferas quod scripseris & emēdes ad exēplaria ea de quib⁹ scripseris diligenter. Et hoc adiurationis genus transcribas & transferas in eum codicem quem descripseris.

Chronica Eusebii Hieronymi Incipit. Prefatio Hieronymi

EVsebius Hieronymus Vincentio & Galieno suis Salutem. Vetus iste disertorum mos fuit ut exercendi ingenii causa grecos libros latino sermone absoluerēt. Et quod plus ī se difficultatis habet poemata illustriū uirorū addita metri necessitate transferrēt. Vnde & noster Tulius Platonis integros libros ad uerbū interpretatus est. Et cū Aratū iā Romanū hexametris uersib⁹ edidisset in xenophontis economico lusit. In quo opere ita sępe aureū illud flumē eloquētię quibusdā scabris & turbulētis obicib⁹ retardat ut qui interpretata nesciūt a Cicerone dicta nō credāt. Difficile est enī alienas linguas insequentē nō alicubi excidere arduū: ut quę in aliena lingua bene dicta sunt: eundē decorē in translatione cōseruēt. Significatū est aliquid unius uerbi proprietate nō habeo meū quo id efficiā: & dum quęro implere sententiā longo ambitu

a 2

The opening page of a 15th century edition of Eusebius' *World History*.
Picture courtesy of Durham University

Keep in touch

To receive updates as the system is developed and made available, get in touch. The *Living Poets* team would like to hear from libraries and archives interested in the OCR Latin software, especially trialling, and customisation; this will help to prove that our 'proof of concept' is working! Email Nick White, IT Research Consultant at: nick.white@durham.ac.uk or for more information about the Living Poets project, visit: livingpoets.dur.ac.uk

Digitise your early book collection

Offering a high degree of accuracy, the Latin OCR system will be made available for libraries and archives to use under a free, open source licence. This will give libraries, archives and publishers a way to transform early printed books in Latin into fully searchable digital documents, free of charge. The programme will be downloadable from the website livingpoets.dur.ac.uk

As part of the project, the research team will be also setting up a not-for-profit company to provide additional customisation services. The team will offer a flexible and cost-effective service to tailor the Latin OCR system for specific collections of books. This will achieve an even higher level of accuracy, with 95-98 per cent of words accurately recognised digitally, and a further manual checking for 100 per cent accuracy.

References

1 Hitchcock, T. 'Confronting the digital: or how academic history writing lost the plot', *Cultural and Social History*, 10(1) 2013, pp. 9-23.

2 Text Creation Partnership. 'Why Keying?' 2014. www.textcreationpartnership.org/why-keying/

3 White, N. 'Training Tesseract for ancient Greek OCR', *Eutypon*, 2012, 28-29, pp. 1-11.